Terminology localization guidelines for the national scenario

Juris Borzovs[1], Ilze Ilziņa[1], Iveta Keiša[2], Mārcis Pinnis[2], Andrejs Vasiļjevs[2]
[1]University of Latvia, [1]Raiņa bulvaris 19, Riga, Latvia
[2]Tilde, Vienibas gatve 75a, Riga, Latvia
E-mail: juris.borzovs@lu.lv, ilze.ilzina@lumii.lv, iveta.keisa@tilde.lv, marcis.pinnis@tilde.lv; andrejs@tilde.com

Abstract

This paper presents a set of principles and practical guidelines for terminology work in the national scenario to ensure a harmonized approach in term localization. These linguistic principles and guidelines are elaborated by the Terminology Commission in Latvia in the domain of Information and Communication Technology (ICT). We also present a novel approach in a corpus-based selection and an evaluation of the most frequently used terms. Analysis of the terms proves that, in general, in the normative terminology work in Latvia localized terms are coined according to these guidelines. We further evaluate how terms included in the database of official terminology are adopted in the general use such as newspaper articles, blogs, forums, websites etc. Our evaluation shows that in a non-normative context the official terminology faces a strong competition from other variations of localized terms. Conclusions and recommendations from lexical analysis of localized terms are provided. We hope that presented guidelines and approach in evaluation will be useful to terminology institutions, regulative authorities and researchers in different countries that are involved in the national terminology work.

Keywords: terminology work, terminology resources, term evaluation, corpus analysis

1. Introduction

This paper presents a set of principles and practical guidelines for terminology work in the national scenario. As described by Henriksen et al. 2006, the national scenario of terminology work deals with the harmonization of national terminology in a well-established infrastructure. It is usually performed by an institution with assigned authority and some regulatory power.

These guidelines have been introduced at and elaborated by the Terminology Commission of the Academy of Sciences of Latvia (LAS-TC) in the domain of Information and Communication Technology (ICT) terminology.

For the most part, ICT terms are created in English and then localized in other languages. By localization of a term we mean the coining of a corresponding term equivalent in the local language to the given English term. Due to the differences in the morphological and terminology traditions in various languages, this localization tends to be rather chaotic. We have developed a set of principles and guidelines described in this paper to ensure a harmonized approach.

We also present our approach in a corpus-based selection and an evaluation of the most frequently used terms. This novel approach allows us to analyse how harmonised terms included in the database of official terminology are adopted in general use.

2. Linguistic principles in the term localization

A newly created ICT term must correspond to the same requirements found at the basis of official terminology. These are: systematicity, precision in meaning, formal brevity, unambiguity, mononymity, contextual independence, and emotional neutrality (Skujiņa & Ilziņa 2011).

English terms don't always fulfil these requirements, therefore difficulties arise when developing corresponding Latvian equivalents. The development of terminology is also hindered by the fact that, in English, ICT terminology doesn't draw a strict distinction between a technical term and professional conversational speech. Likewise, in the choice of terms, the requirements put forth for a technical term are not observed, in the traditional understanding of terminology (Borzovs & Ilziņa 2010).

The creators of ICT terms are also faced with the large number of metaphors found in English terminology. This is a common problem in terminology, even with ISO standards, and Latvian ICT terms have been unable to avoid these metaphors.

Newly localized ICT terms can be clustered into three categories:

1) Terms that are created based on words used in everyday language and included in general lexicons.
2) Terms that are coined by such borrowings from other languages that are already used in the local language.
3) Neologisms that can be either completely new words rooted in the local language or translingual borrowings transferred and adapted from the original language (Borzovs & Ilziņa 2010).

In the localization of terms, a number of specific principles of correspondence should be observed (Borzovs et al. 2002):

- Semantic correspondence principle;
- Formal correspondence principle;
- Functional correspondence principle;
- Term dissemination principle;
- Tradition principle.

The **semantic correspondence** principle holds that, when creating terms, each lexical pattern has a specific semantic weight that is characteristic of the corresponding language system.

The **formal correspondence** principle holds that words that share a similar form in the original language should share a similar form in the target language. New forms, new

words, and syntactical units are developed based on stable models.

The **functional correspondence** principle is related to such basic signs as the brevity of a term, the ease of use, and euphoniousness.

This principle also holds that short terms are easy to use: they form a system more easily and new elements can be added to them, thus creating sub-concept terms.

By borrowing terms or creating a new term, attention is usually paid to the ease of use, that is, making sure a term can be easily conjugated and easily use in combinations of words.

Euphoniousness is important both when borrowing a term from other languages, and when creating a neologism. In Latvian, the localizers of ICT terms allow for the borrowing of terms from other languages, though always paying special attention to the euphoniousness of the language.

In order to observe the **term dissemination principle**, LAS-TC pays special attention to terms that could be a part of everyday use, that is, used widely and often. They should be short, concise, and euphonious, and must conform to all the criteria.

The **tradition principle** applies if a term is already widely used, or if it was confirmed several years ago.

The goal of terminology – and of Latvian ICT term localizers – is to make communication more effective. In each sector, the process of developing terms should be based on the experience of terminology work, using the existing system and developing it with principles formulated during practical terminology work.

3. Practical guidelines for term localization

When localizing ICT terms into Latvian, the following ten guidelines are created and observed in the localization process:

1. One term in the original language should correspond to one specific term in the target language.
2. Differing terms in the original language should be given differing terms in the target language.
3. If a term is ambiguous in the original language, a word with a similar range of ambiguity should be chosen in the target language.
4. When coining a neologism, observe its suitability in the corresponding term system and similarity with related and analogic terms.
5. One should choose a term's equivalent so that, when translating it back to the original language, the same original word is the clear choice.
6. When borrowing a word, pay heed to how well it fits into the target language semantically, phonetically, and morphologically.
7. When faced with a choice between international borrowings and native words, preference is given to native words.
8. Do not change, without a sound basis, a word already used in practice.
9. More attention should be paid to words widely used by the general public. They should be short, precise, euphonious, and easy to understand.
10. None of the aforementioned principles shall be made absolute.

4. Corpus based approach in identification of the most frequent terms

For the evaluation of localized ICT terminology consisting of more than 7,000 term entries, we wanted to identify the most frequent terms.

In practice, a single concept in Latvian can be denoted with many different terms. These are not only official terms standardized by the LAS-TC, but also other forms widely used in the public sphere and in informal communication. For instance, the concept *computer* can be translated in Latvian as *dators*, *kompjūters*, or *skaitļotājs*. English, on the other hand, is less ambiguous and a single concept is usually denoted with one lexical equivalent.

Therefore, in order to identify the most frequent ICT terms in Latvian, we used an English-Latvian bilingual corpus that was automatically collected from the Web. In the collection process, only Web sites containing both Latvian and English content were crawled. The statistics of collected bilingual corpora are given in Table 1.

Further, we calculated the statistics of the English terms in the bilingual corpus. When calculating the term occurrence counts, different surface forms of a single term were grouped together. The term surface forms were lowercased and stemmed with the Porter Stemmer (Robertson et al. 1980). Then we selected the 200 most frequent English terms that have equivalents in Latvian which have been officially approved by LAS-TC.

Parameter	English	Latvian
Sentences	3 358 914	3 404 515
Tokens	44 482 878	44 613 452
Unique sentences	2 877 176	2 906 786
Tokens in unique sentences	38 713 499	38 763 916

Table 1: Statistics of the English-Latvian bilingual Web corpus

For term translation equivalent lookup, we used the termnet.lv (Skadiņš, Vasiļjevs 2004) termbase, which provides access to the official ICT term collection of the Information Technology, Telecommunications and Electronics Sub-Commission of the Terminology Commission of the Latvian Academy of Sciences (LAS-TC-ITTE). The collection contains more than 7,200 ICT term entries. Because new terms require some time to get into public circulation, we excluded from the analysis all terms that were adopted after 2011. As a result, 6,886 terms were used in the further analysis process. The top 10

English ICT terms from the English-Latvian bilingual corpus are given in Table 2.

Term	Frequency	Term	Frequency
mode	82 163	file	32 005
warning	79 369	service	29 672
window	62 512	download	26 124
click	37 673	information	26 010
key	34 482	data	19 967

Table 2: Ten most frequently used English ICT terms found in the English-Latvian bilingual corpus

The set of the 200 most frequent English terms included also such common ICT terms as: *system, help, search, user, computer, internet, location, security, program, web, message, link, code, form, online, software, folder, network, application, field, comment, server, control, guest, format, table, bit, card, PC, display, menu, address, button*, and others.

The next task was to identify which Latvian terms are used in practice as translation equivalents of the English ICT terms. We performed this task in a semi-automatic manner using three approaches:

1. At first we identified all Latvian equivalents of the English Terms using the official term collection approved by the LAS-TC-ITTE.
2. In order to identify non-official term equivalents we used the bilingual English-Latvian corpus. The corpus was first aligned at the phrasal level using the statistical machine translation platform LetsMT [1] (Vasiļjevs et al. 2010). The LetsMT platform is based on the Moses SMT Toolkit (Koehn et al., 2007), which performs word alignment at the sentence level and then extracts bilingual phrases in the form of a Moses phrase table. We used the Moses phrase table as a term translation equivalent lookup table in order to acquire for each English term translation equivalents in Latvian (including different surface forms). As the automatic alignment creates noise, a field expert manually revised the term pairs and removed all wrongly aligned term translation equivalents.
3. Although the bilingual corpus is relatively large, it does not contain all term translation equivalents that are used in public communication. Therefore, two field experts manually revised the results and added additional colloquial term equivalents.

As a result of the semi-automatic process for the 200 most frequently used English terms, we identified 997 different term translation equivalents in Latvian (excluding surface forms). This means that, on average, each English term had five translation equivalents in Latvian.

[1] LetsMT platform is accessible online: *http://www.letsmt.eu*.

Next we calculated the frequency of Latvian term equivalents using an only ICT related text monolingual corpus, which was also collected from the Web. For corpora collection we used the FMC tools - the Focussed Monolingual Crawler (Mastropavlos and Papavassiliou, 2011). The corpus consists of news articles (*Apollo.lv, Delfi.lv, Diena.lv*, etc.), blog posts (*krizdabz.lv, aidzis.lv, knagis.miga.lv*, etc.), product reviews (*kakao.lv, androids.lv, iPods.lv*, etc.), and press releases and documentation articles (*microsoft.lv, samsung.lv, lattelecom.lv*, etc.) that were downloaded from Web domains directly related to ICT or containing articles on different ICT related topics. The corpus statistics are given in Table 3.

Parameter	Latvian
Web domains (and specialised subdomains)	204
Unique documents	15 007
Sentences	2 275 019
Tokens	14 558 150
Unique sentences	434 664
Tokens in unique sentences	5 485 361

Table 3: Latvian ICT Web corpus statistics

As the Web corpus contains also static page content (for instance, menu texts, copyright information, reappearing advertising, etc.), we performed corpus filtering by extracting only unique paragraphs from all pages belonging to a single Web domain. The filtering is applied using Web domains as a grouping criteria in order to account also static content that is frequently re-used in multiple Web domains and can potentially contain important to the ICT field terms. Because some documents within one domain contained redundant information, the final number of productive documents (containing at least one unique paragraph within the document's Web domain) was reduced to 9 979. The top 10 domains in respect to the number of productive documents are listed in Table 4.

Once the corpus was collected and filtered, we calculated the occurrence statistics of the 997 distinct Latvian term variants within the corpus. We calculated the statistics of every surface form of a term in a given Web domain and aggregated the counts for every surface form, for every term in Latvian and also for every translation equivalent in English (that we acquired using the semi-automatically created term glossary).

Web domain	Productive documents
datuve.lv	803
kakao.lv	713
parasts.lv	585
krizdabz.lv	582
androids.lv	520
latfoto.lv	508
ipods.lv	416
forums.lattelecom.lv	391
samsung.com	349
lattelecom.lv	348

Table 4: The top 10 domains of the ICT related text corpus

An excerpt of the aggregated results in a pivot table for the English term *mode* and its Latvian equivalent *režīms* is given in Figure 1. The figure shows the occurrence count of the term *mode* in the bilingual corpus (82 163), the sum of its equivalent occurrences in the Latvian focussed corpus (2 387), a Latvian equivalent in a form where the ending is dropped for aggregation purposes (*režīm* and the respective occurrence count – 1 898), a surface form in Latvian (*režīmā* and the respective occurrence count – 797) and a list of Web domains where the surface form occurred sorted in a descending order depending on the occurrence counts in the respective Web domains.

The acquired terms and the term usage statistics were further used in a manual process in order to analyse the official ICT terminology usage trends in Latvian.

Row Labels	Sum of LV_COUNT
82163	2387
mode	2387
režīm	1898
režīmā	797
krizdabz.lv	137
latfoto.lv	117
androids.lv	85
datuve.lv	59
kakao.lv	48
macpasaule.lv	33
gamez.lv	31
zparks.lv	29
windows.microsoft.com	28

Figure 1: An excerpt of the aggregated results of the term frequency analysis in the ICT related text corpus

5. Evaluation of localized terminology

We performed a manual evaluation of the terminology data, prepared as described in the previous section – the most frequent English terms, Latvian equivalents from the official database, and Latvian terms extracted from the Web with their usage statistics.

For the 200 most frequent English ICT terms, we identified 281 Latvian equivalents in the official terminology database. These terms were clustered into the categories defined in the Section 2:

- 115 terms were coined from native language words that are part of the general lexicon (term category 1)
- 104 terms were coined from international borrowings adapted in Latvian before the advent of the computer era (term category 2)
- 62 terms were neologisms (term category 3), including:
  - 39 neologisms rooted in the local language patterns;
  - 19 neologisms created by phonetic transliteration of the original term;
  - 4 terms created by transcription of the original term.

An analysis of the terms shows that, in general, the guidelines described in Section 2 were followed in the creation of the official terminology in Latvia.

A second task was to evaluate the adaptation rate of the official terms in general use of public communication. We compared how frequently the official terms are used in comparison to other Latvian translation equivalents for the same English term (e.g., usage of the official term *dators* compared to *kompjūters*, *skaitļotājs*, and other forms not recommended by LAS-TC).

We counted the total number of occurrences of all Latvian terms that are equivalents of the same English term and calculated their relative frequencies using the following formula:

$$Relative\ Frequency\ (term_i) = \frac{Count(term_i)}{\sum_{term_j \in Y} Count(term_j)} \times 100\% \quad (1)$$

where $term_i$ and $term_j$ are the i^{th} and j^{th} Latvian translation equivalent of an English term, and Y is the set of all Latvian translation equivalents of the particular English term.

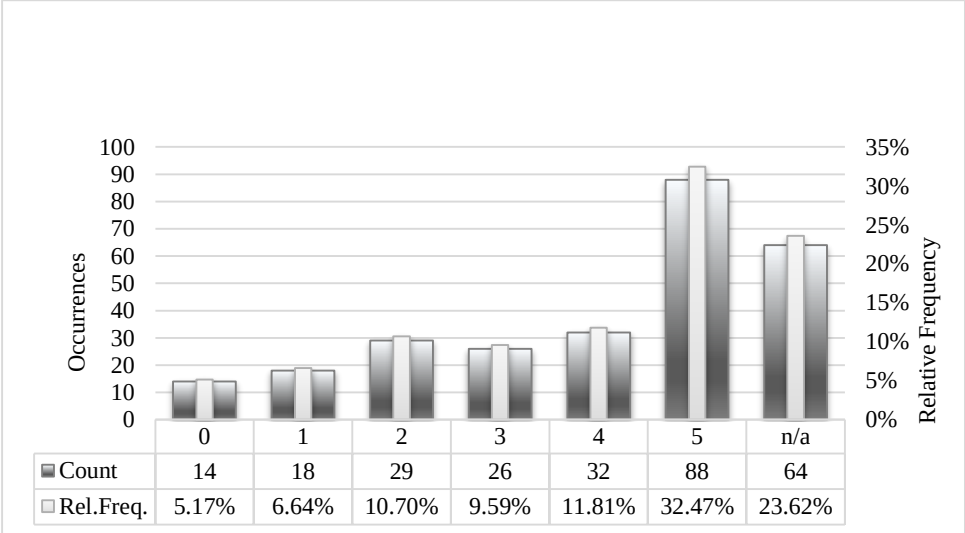
Based on the relative frequencies we assigned a usage grade on a scale of 1 to 5 for each of the Latvian translation equivalents of the English terms:

0-1% - grade 0 (not used)
1-10% - grade 1 (rarely used)
11-30% - grade 2 (occasionally used)
31-50% - grade 3 (second choice)
51-80% - grade 4 (preferred)
81-100% - grade 5 (fully adopted)

The terms that were too ambiguous to distinguish in the analysis from general language words (e.g., Latvian equivalents for *set, map, sign, etc.*) were excluded from this analysis. In total we excluded 24% of Latvian terms because of the lexical ambiguity (n/a in the Figure 1).

By analysing terms in every grade we can draw several conclusions outlined below:

- Popular abbreviations are almost always preferred over the full terms (e.g., *PC* instead of *personālais dators* for the English *personal computer*);
- Official neologisms are rarely used if there are common words with similar meaning (e.g., *serviss* instead of *pakalpe* for *service*);
- Transcribed borrowings are rarely used if there are common words with a similar meaning;
- For terms metaphoric in English, users prefer similar metaphors in Latvian instead of neologisms;



	0	1	2	3	4	5	n/a
Count	14	18	29	26	32	88	64
Rel.Freq.	5.17%	6.64%	10.70%	9.59%	11.81%	32.47%	23.62%

Figure 2: Adaptation grade of the official Latvian terminology

- Longer terms are an occasional choice, giving preference to shorter forms (e.g., *ziņa* instead of *ziņojums* for *message*, *pārlūks* instead of *pārlūkprogramma* for *browser*);
- A native language term is an occasional choice if there is a traditional international borrowing as an alternative (e.g., *digitāls* instead of *ciparu* for *digital*);
- A well-formed neologism can become a second choice if a borrowing has already been in use before;
- Users easily accept common words that have been assigned a new meaning when used as terms.

6. Conclusions

We described approach, linguistic principles and practical guidelines elaborated and applied by the regulative terminology authority in Latvia. We further presented a novel corpus based approach in evaluation of the normative terms selecting the most frequent terms in ICT domain. Analysis of the terms shows that, in general, the guidelines were followed in the creation of the official terminology in Latvia. Further analysis showed that in the non-normative context official terminology faces a strong competition from other variants of term localizations. Still our analysis proves that official terminology is more widely used.

We hope that our approach and experience will be useful to institutions in different countries that are involved in national terminology work.

7. Acknowledgements

The research leading to these results has received funding from the research project "Optimization methods of large scale statistical models for innovative machine translation technologies" of European Regional Development Fund, contract nr. 2013/0038/2DP/2.1.1.1.0/13/APIA/VIAA/029.

8. References

Borzovs, J., I. Ilziņa and I. Vancāne. (2002). „The Main Guidelines of Creating IT&T Terminology." *Terminology and Technology Transfer in the Multiligual Information Society. 2nd International Conference on Terminology in Commemoration of E. Drezen's 110th anniversary*. Pp. 25 – 32. Riga, Latvia.

Borzovs, J., I. Ilziņa. (2010). „The Problems of Latvian ICT Terms and English Borrowings." *Leksikografija ir leksikoloģija.* Lietuviu Kalbos institutas, pp. 329 – 340. Vilnius, Lithuania.

Henriksen, L., Povlsen, C., & Vasiljevs, A. (2006). EuroTermBank–a Terminology Resource based on Best Practice. In *Proceedings of LREC 2006, the 5th International Conference on Language Resources and Evaluation*, pp. 243–246, Genoa.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., … Herbst, E. (2007). Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pp. 177–180). Stroudsburg, PA, USA: Association for Computational Linguistics.

Mastropavlos, N., & Papavassiliou, V. (2011). Automatic acquisition of bilingual language resources. In *Proceedings of the 10th International Conference of Greek Linguistics, Komotini, Greece*.

Robertson, S. E., van Rijsbergen, C. J., & Porter, M. F. (1980). Probabilistic models of indexing and searching. In *Proceedings of the 3rd annual ACM conference on Research and development in information retrieval*, pp. 35-56. Butterworth & Co..

Skujiņa, V., I. Ilziņa. (2011). „The Development of Latvian Terminology under the Impact of Translation." *Terminologija*, no. 18. Lietuviu Kalbos institutas, pp. 43 – 50. Vilnius, Lithuania.

Vasiļjevs A., Gornostay T., Skadiņš R. (2010) LetsMT! – Online Platform for Sharing Training Data and Building User Tailored Machine Translation. In *Proceedings of the 4th International Conference "Human Language Technologies – The Baltic Perspective"*, October, 2010, Riga, Latvia.